DOCUMENT RESUME

ED 286 913 TM 870 545

AUTHOR Stanley, Julian C.
TITLE Gender Differences on the College Board Achievement Tests and the Advanced Placement Examinations: Effect Sizes Versus Some Upper-Tail Ratios.
PUB DATE 24 Apr 87
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Advanced Placement; *College Entrance Examinations; *Effect Size; Equivalency Tests; Higher Education; High Schools; Intellectual Disciplines; *Sex Differences
IDENTIFIERS *Advanced Placement Examinations (CEEB); *College Board Achievement Tests

ABSTRACT

Gender differences were determined for scores on the 14 College Entrance Examination Board achievement tests taken between 1982 and 1985. Effect sizes showed a tendency for males to score higher relative to females, while the percentage of male examinees increased. Based on Cohen's classification of effect sizes, differences in six of the 14 tests were small and two were moderate. All favored males. At the upper end of the test score scale, physics in 1983 showed the largest effect size, 0.060. Classified as moderate by Cohen, this effect size corresponded to almost 14 males to each female who scored 700-800, perhaps meriting a stronger classification than moderate. Other examples include scores for European history, Latin, biology, English composition, and English literature. For Advanced Placement examination scores between 1984 and 1986, the largest mean effect size was for computer science, taken by 85% males. Four other science examinations--three physics tests and chemistry--showed the next largest effect sizes and were also taken by the most males. These differences were not as strong as those seen on the College Board achievement tests. (MGD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286913

Gender Differences on the College Board Achievement Tests and the Advanced Placement Examinations: Effect Sizes Versus Some Upper-Tail Ratios*

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. C. Stanley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Julian C. Stanley
Study of Mathematically Precocious Youth
Johns Hopkins University
Baltimore, MD 21218

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The College Board achievement tests are offered several times each school year so that applicants to selective colleges may take the ones (usually three) they need. Each test consists of a large number of multiple-choice items to be answered in sixty minutes. Three tests may be taken on a single morning. All the tests are high-school level, not to be confused with the College Board's Advanced Placement Program examinations. The latter are meant to cover the equivalent of two semesters of course work in the subject at a selective college or university--that is, they are college-freshman level.

The Achievement Tests

During each of the years 1982, 1983, 1984, and 1985 males and females

---

*Paper read at the annual meeting of the American Educational Research Association in Washington, D.C., on Friday, 24 April 1987, during the 8:15-10:15 a.m. symposium entitled "Sex Differences in Cognitive Abilities and Achievement" in the Diplomat Room, West Floor of the Omni Shoreham Hotel. I thank Linda E. Brody, Susan L. Dauber, Robert A. Gordon, Ann E. Lupkowski, Barbara S. K. Stanley, and Howard Wainer for helpful suggestions.

TM 870 645

took one or more of 14 tests. The number of examinees ranged from 144 males taking the Hebrew test in 1984 to 95,329 females taking the English Composition test in 1985. Most numbers were over 1000. For each year the Profiles, College-Bound Senior volume compiled by Leonard Ramist and Solomon Arbeiter[1] and published by the Admissions Testing Program of The College Board gives rather complete statistics about the scores earned and background characteristics of the examinees. For the present study the Ns, means,and standard deviations for each sex were used to compute the percentage of examinees who are male. The latter enables one to make standardized comparisons across tests. As will be illustrated later, however, such measures tend to underestimate gender differences in the upper part of test-score distributions.

For all the papers being presented in this symposium, effect size is defined as the mean of the males minus the mean of the females divided by the positive square root of the mean of the two variances. This procedure gives equal weight to the standard deviation of each sex on the test. It does not use the pooling of sums of squares involved in Student's t-test, because that method allows the N of the larger group of examinees to weight that sex's variance more than the other's. Where the two variances are not very different, the two procedures yield similar results.

Figure 1 shows on its horizontal axis the median percent of males taking each test, determined from the four yearly percents. On the vertical axis is shown the mean effect size, which is the male minus female mean difference expressed in standard-deviation units; the four yearly effect sizes for each test were averaged.[2]

---

[1] By only Arbeiter in 1984.

[2] Medians were used for the percentages because they cannot be considered to form an interval scale, whereas the effect sizes can.

For example, 28.5 percent of the examinees taking the French test were male, so of course 71.5 percent were female. The mean effect size for French was only .08, slightly favoring the males. Of those taking the literature test, 35 percent were male. Its mean effect size, -.09, slightly favored the females. English Composition, taken by almost exactly the same percent of males and females, favored females most. Its effect size, -.10, was small, however. Effect sizes for Hebrew (-.001) and German (-.006) also favored females but were essentially zero.

Figure 1 reveals a strong tendency for males to score higher and higher, relative to the females, as the percentage of examinees who are male increases. At 65 percent and effect size .58, European History is almost at the level of physics, 81 percent and .59. Sex differences for Biology, Level 1 Mathematics, American History, Level 2 Mathematics, and Chemistry are similar, their mean effect sizes ranging from .36 to .40.

Cohen (1977) arbitrarily classifies effect sizes derived from mean differences as negligible if below .20, "small" if between .20 and .50, "moderate" size if between .50 and .80, and "large" if .80 or more. By these criteria, 6 of the 14 differences are "small" and 2 are "moderate." All these 8 favor males. Using Cohen's classifications, one concludes that distributions of the scores of the two sexes overlap considerably.

If one is interested in selection at the upper end of the test-score scale, rather than in overlap of distributions, what do the effect sizes imply? Let us start with the largest effect size for Physics during the four years, .60 in 1983. This value lies below the middle of Cohen's "moderate" category. How many males and females scored 700-800 on the test, that being the range usually favored by admissions committees of highly select colleges? Of the 3163 females tested, 192 did. Of the 13,344 males, 2668 did. Thus,

nearly 14 males earned 700 or more for each female who did so. The ratio of percentage of males scoring 700-800 to percentage of females scoring in that range was 3.3 to 1. That large discrepancy seems to deserve a stronger appelation than "moderate."

The largest of all the 56 effect sizes was .63 for European History in 1985. Of the 2053 males who took that test then, 216 scored 700-800. Of the 1077 females, 22 did. Thus, almost 10 males were in that range for each female who was. The ratio of percentages was 5.2, again much larger than one might expect for an effect size lying below the middle of Cohen's "moderate" category. This illustrates why Cohen warns his readers not to stop with effect sizes, but instead to probe the data further.

Let us try an effect size that lies in the upper part of the negligible category, .18 for Latin in 1982. Its percentage ratio was 1.5 to 1, a sizable difference.

The smallest Biology effect size was .34 in 1985. This is the middle of the "small" category, but the percentage ratio was 1.85:1, hardly small.

The highest of the 8 effect sizes favoring the females on English Composition or Literature was .12 in 1984. The scores underlying this yield a percentage ratio of 1.18 females for each 1 male.

It is clear that, relative to the males, the females fared poorly on many of the 14 College Board high-school achievement tests and not excellently on any of them. Are these particular achievement tests in some ways "unfair" to females, or do females tend to score lower than males on most standardized mental aptitude and achievement tests? Let us compare these results with analogous figures for the College Board Advanced Placement Program's college-level tests, which students are not required to take in order to apply for admission to the nation's most select colleges.

Gender Differences on the 24 AP Examinations

In 1984, 1985, and 1986 interested males and females of any age could take one or more of the 24 AP exams in May of the year. Numbers doing so ranged from 94 males on Music: Listening & Literature in 1984 to 44,436 females on English Literature and Composition in 1986. In 1986, only one of the 24 tests had fewer than 256 male and 284 female examinees, so effect sizes from year to year tend to be fairly stable.

In Figure 2 is plotted, on the vertical axis, the mean effect size on each of the 24 AP examinations for the three-year period 1984-86, against the median percent males on the horizontal axis. The largest mean effect size (.50) is for computer science, taken by 85 percent males. The five largest effect sizes in the graph are on exams taken by the most males: .29 and 86 percent for Physics C: Electricity and Magnetism, Computer Science as noted above, .37 and 85 percent for Physics C: Mechanics, .41 and 77 percent for Physics B, and .33 and 70 percent for Chemistry. All of these are science subjects.

Only one, Computer Science, qualifies (barely) for Cohen's "moderate" category. As we have seen earlier, however, his categories tend greatly to underestimate the selective ratio of the test for high scorers. A look at Table 1 reveals how potent a gender-related selector the Computer Science exam is for getting credit at a college or university where only 4s and 5s qualify. The good news in the table is that the effect size favoring males dropped greatly in 1986, from .59 in 1984 and .55 in 1985 to .37 in 1986. May the decline continue! The bad news is that the percentage of males taking the test remained constant and huge, 85 for 1984, 1985, and 1986.

There are two main ways to look at the numbers of males and females scoring 4s or 5s on an AP examination. One is to take the ratio of number of

males earning those scores in a given year to the number of females who do. For Computer Science these yield ratios of 12:1 for 1984 and 1985, and 9:1 for 1986. Another way is to take the percent of males earning 4s or 5s in a given year divided by the percent of females who do. These Computer Science ratios are 2.1:1, 2.1:1, and 1.6:1, respectively. Either of these methods reveals discrepancies that, in my opinion, should be food for thought. The AP Computer Science exam, offered only from 1984 onward, shows females lagging greatly in the skills tested by it. Because, as data in this paper and other papers in this symposium demonstrate, females also tend to score considerably less than males on other key aspects of scientific background and aptitude (e.g., mechanical comprehension, mathematical reasoning, physics, and chemistry), many seem likely to find themselves unable to compete effectively in the physical and mathematical sciences and engineering. Doesn't this warrant far more attention than just the study of item types? The problem, whatever its basis, seems much more complex than that.

Five other examinations yielded effect sizes in the .20s, which Cohen calls "small." These were Music: Listening & Literature, .27; for European History, .26 (vs. the .58 found for the analogous high-school achievement test); American History, .24 (vs. .40); Biology, .22 (vs. .36), and Calculus AB, .20 (vs. .18 for Calculus BC, the more advanced level).

The largest of the 7 mean effect sizes favoring females was .15, for Latin: Catullus & Horace, taken by 55 percent males. Over the three-year period, however, the males had been catching up on this, with effect sizes favoring females (i.e., those having negative signs) declining as follows: -.32, -.08, and -.06. Females had gained on American History (.26, .25, .22), Art Studio: General, (.18, .15, -.08), Computer Science (as already noted), French Language (.08, .06, .04), Music Theory (.15, .03,

-.04), and Physics B (.47, .39, .38).

Males rose steadily on Biology (.21, .22, .24), Chemistry (.31, .31, .37), English Literature & Composition (-.10, -.04, -.01), German Language (-.19, -.12, -.10), and (as noted above) Latin: Catullus & Horace. The number of up or down "runs" is statistically greater than chance: 3 of 14 achievement tests (chance probability of a run is 1/24) and 11 of the 24 AP examinations (probability of a run is 1/6). Some sizable fluctuations will occur from year to year, especially on the low-enrollment examinations. Whether the apparently steady gains or losses of either sex will continue in subsequent years remains to be seen.

Let us examine what an effect size in the middle of the "small" range connotes for earning 4s and 5s on AP exams. Of all 72 effect sizes, the closest one to .35 is .357 for Physics C: Mechanics in 1986. For every female who scored 4 or 5, 7.7 males did. The ratio of the percent males scoring that high (55) to the percent females doing so (39) is 1.42 to 1. A 42% excess of one gender over the other hardly seems "small."

Finally, let us examine the effect size nearest .20, but just less than it, .1987 on Calculus BC in 1984. The needed ratios prove to be 2.3:1 and 1.2:1, respectively. The latter means that (more precisely) 21.6% more males than females earned 4s and 5s. Is that "negligible"?

Summary

Gender differences tend to be somewhat stronger on the required-for-college high school achievement tests than the optional AP examinations. Nevertheless, results of the two analyses agree substantially. They should prompt searching for ways to lessen the discrepancies, especially in science, mathematical sciences, and history. Doing away with standardized achievement testing does not seem to me a reasonable "solution."

A further caution is needed. Recently, several persons have proposed that for college admissions the SAT be replaced by College Board achievement tests. Even a casual comparison of my Figure 1 with Susan Dauber's graph concerning SAT suggests that females might not be served well by such a swap.

References

Advanced Placement Program. (1984). National summary reports, 1984. New York: College Entrance Examination Board. Pp. 4-5.

Advanced Placement Program. (1985). National summary reports, 1985. New York: College Entrance Examination Board. Pp. 4-5.

Advanced Placement Program. (1986). National summary reports, 1986. New York: College Entrance Examination Board. Pp. 4-5.

Arbeiter, S. (1984). Profiles, college-bound seniors, 1984. New York: College Entrance Examination Board. Pp. 13-14 and 23-24.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Ramist, L., & Arbeiter, A. (1984). Profiles, college-bound seniors, 1982. New York: College Entrance Examination Board. Pp. 13-14 and 23-24.

Ramist, L., & Arbeiter, A. (1984). Profiles, college-bound seniors, 1983. New York: College Entrance Examination Board. Pp. 13-14 and 23-24.

Ramist, L., & Arbeiter, S. (1986). Profiles, college-bound seniors, 1985. New York: College Entrance Examination Board. Pp. 13-14 and 23-24.

Figure 1: Average Standardized Gender Differences on Each of the 14 College Board Achievement Tests for the Years 1982-85, Plotted Against the Median Percent Males Who Took the Test

(Parentheses enclosing effect sizes designate the 4 tests on which females scored higher than males.)

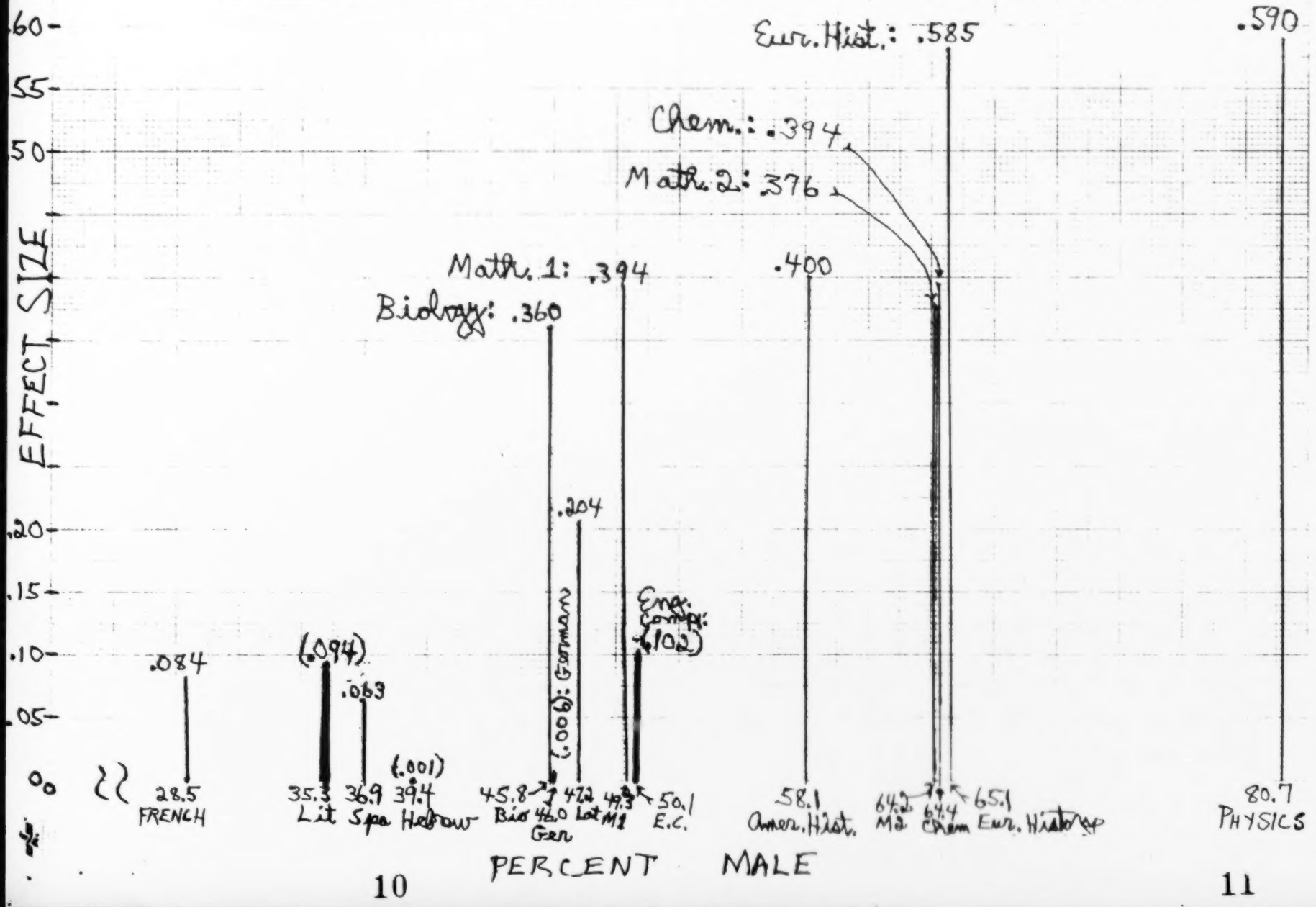

Table 2

Sex Differences on AP Computer Science Examination

Year	Mean for Males	Mean for Females	S.D. for Males	S.D. for Females	Ratio of Males to Females	Ratio of Males to Females Scoring 4s or 5s	Ratio of the Proportion of Males Scoring 4s or 5s to the Proportion of Females Scoring 4s or 5s	Effect Size
986	3.06	2.58	1.34	1.25	6819:1212 = 5.63:1	8.82:1	1.57:1	0.37
985	3.07	2.36	1.33	1.24	5857:1026 = 5.71:1	12.2:1	2.13:1	0.55
984	3.19	2.43	1.32	1.28	3562:623 = 5.72:1	11.9:1	2.08:1	0.59

Table 1

Gender Differences on 24 Advanced Placement Program Examinations

Humanities			STANDARDIZED EFFECTS Male - Female	Sciences
Latin CH		Ger. Lang.	-.15	
		Fr. Lit.	-.10	
Latin Ver	Eng. Lit	Sp. Lang., Sp. Lit.	-.05	
		Art Hist.	-.00	
Music Th	Eng. Lang.	Fr. Lang.	.05	
	Art SD	Art SG	.10	
			.15	
			.20	Biology, Calc AB, Calc BC
Music L&L	Am. Hist.	Eur. Hist.	.25	
			.30	Physics E&M
			.35	Chem, Physics - Mech.
			.40	Physics B
			.45	
			.50	Comp. Sci

Prepared by that master of graphic presentation, Dr. Howard Wainer of the Educational Testing Service.

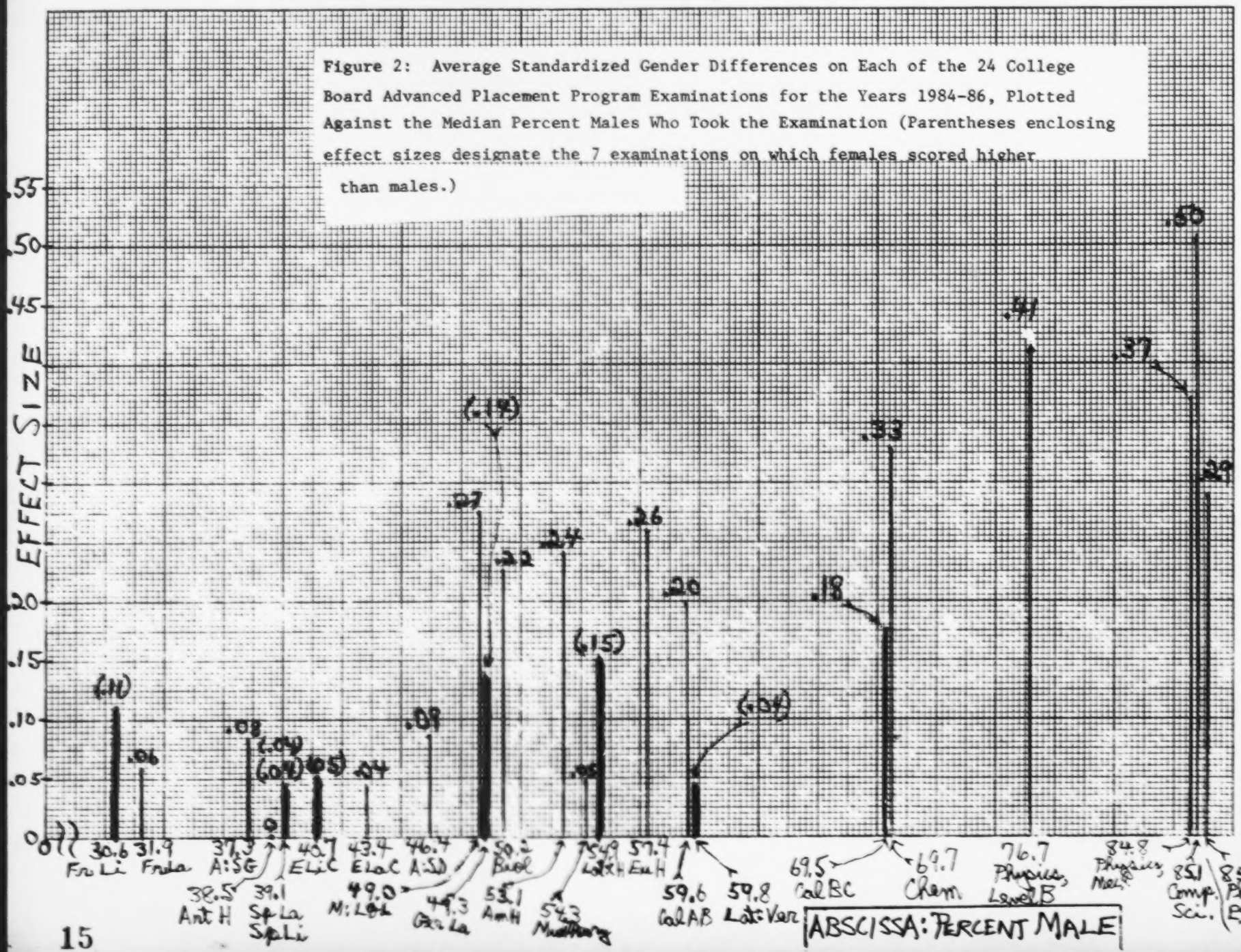

Figure 2: Average Standardized Gender Differences on Each of the 24 College Board Advanced Placement Program Examinations for the Years 1984-86, Plotted Against the Median Percent Males Who Took the Examination (Parentheses enclosing effect sizes designate the 7 examinations on which females scored higher than males.)